NBI Data Access

These design notes concern exposing the UNO NBI data (

~~NDS-992~~ - Define/prototype how we will incorporate NBI data into Workbench RESOLVED

) to users via workbench.

See also Shared data directories

Background

From Robin Gandhi, Univ. of Nebrask at Omaha:

Compute and query infrastructure for National Bridge Inventory data: Federal Highway Administration (FHWA) requires all state Departments of Transportation/Roads to annually report information on bridges and tunnels that have road traffic. This data, which is called the National Bridge Inventory (NBI), is made available through position aligned or comma separated values, sometimes compressed, on the FHWA website (https://www.fhwa.dot.gov/bridge/nbi/ascii.cfm). Since 1992, this dataset has collected approximately 17 million bridge inspection records. Each bridge inspection record conforms to a data coding guide, which allows the dataset to capture a great amount of information in a dense format. Due to the shear size of these records, simple tools such as Excel are not suitable for any advanced data analytics. This has been noted by many researchers attempting to analyze this dataset. To make this dataset more accessible, we have developed scripts to transfer this dataset into a big data pipeline. In particular, we have setup a MongoDB instance using infrastructure available from a cloud provider (digital ocean). A simple example of the data analytics for all the bridges in Nebraska, which is possible through the new prototype we developed, is available here: http://faculty.ist.unomaha.edu/rgandhi/r/mongoNBI.html. All data export scripts (in active development) are available on Github (https://github.com/kaleoyster/ProjectNBI) to replicate these activities.

Discussions began in June 2017 to "transition" to DataDNS. This project presents a couple of interesting opportunities:

- Hosting data in an active (i.e., database) format. While we can keep a copy of the raw data, the active data, even if read only, is likely of more interest to the community.
- The raw data is available from DOT, but software has been written to convert to a more usable format.
- UNO has provided a sample Jupyter notebook to analyze the data.

Workbench, Share, and DataDNS

- The Labs Workbench service platform supports launching a variety of services, including analysis environments. It would be easy to include the UNO Jupyter notebook in a Workbench instance.
- Workbench could easily host a "public" MongoDB for the NBI data. Of course, this raises questions (how long, how do we maintain it going forward, is there an SLA, do we offer this for everyone?) As a pilot, this seems like an interesting opportunity.
- The NDS Share Globus endpoint provides a place and mechanism to transfer data. Via Globus Publish, it also supports metadata for dataset description.
- We end up with the following:
    - Raw data via Globus on santiago
    - Metadata record in Globus onling
    - Code in Zenodo
    - MongoDB container running in "public" space on Workbench
    - Raw data optionally mounted via NFS (however, Workbench is currently at SDSC)
    - Notebook container available in Workbench
- We could either install Workbench on santiago or put the data on the SDSC Workbench instance while exploring how best to expose it.

Processing steps for NBI Data

Download the NBI data

```
git clone https://github.com/kaleoyster/ProjectNBI
```

```
python ./nbiCsvJsonConverter-2/Downloadv1.py
```

Creates directory NBIDATA containing the raw data

```
python3 ./nbiCsvJsonConverter-2/ProcessMain.py
```

Converts ingests the data into a MongoDB

Data is 6.3 GB uncompressed CSV

MongoDB

Getting the data to Workbench

We can initially add the shared directory to Gluster and transfer or download the data directly.

For this project, the raw data is probably less of a concern than the Mongo data – which poses an interesting question about data sharing.

We could setup a globus endpoint or use Santiago?

- santiago.ncsa.illinois.edu runs a globus endpoint (https://data.share.nationaldataservice.org/)
- The data exposed through santiago could be mounted on Workbench – but it's at NCSA, not SDSC

Import into MongoDB

We will host the raw data under /shared/NBIDATA/ as a read-only volume

We will also host an instance of MongoDB in the "public"? namespace with the official database

Users can access via nbidata.public

This will require running a process to ingest the data.

The "public" namespace can have no service timeouts

Mongo must be accessible to all namespaces, even after we apply network security policies.

Sample Document


```
{
        "_id" : ObjectId("59b8519bf6b8e300bb668a93"),
        "year" : 1992,
        "stateCode" : "02",
        "structureNumber" : "0175",
        "inventoryRoute" : {
                "recordType" : "1",
                "routeSigningPrefix" : -1,
                "designatedLevelOfService" : -1,
                "routeNumber" : "NA",
                "directionalSuffix" : -1
        },
        "highwayAgencyDistrict" : "00",
        "countyCode" : 0,
        "placeCode" : 0,
        "featuresIntersected" : {
                "featuresIntersected" : "NA",
                "criticalFacilityIndicator" : "NA"
        },
        "facilityCarriedByStructure" : "NA",
        "location" : "NA",
        "InventoryRTeMinVertClearance" : 0,
        "kilometerpoint" : -1,
        "baseHighwayPoint" : -1,
        "inventoryRouteSubrouteNumber" : {
                "LRSInventoryRoute" : "NA"
        },
        "latitude" : 0,
        "longitude" : 0,
        "bypassDetourLength" : 0,
        "toll" : -1,
        "maintenanceReponsibility" : -1,
        "owner" : -1,
        "functionalClassOfInventoryRte" : -1,
        "yearBuilt" : -1,
        "lanesOnUnderStructure" : {
                "lanesOnStructure" : -1,
                "lanesUnderStructure" : 0
        },
        "averageDailyTraffic" : 0,
        "yearOfAverageDailyTraffic" : -1,
        "designLoad" : 0,
        "approachRoadwayWidth" : 0,
        "bridgeMedian" : 0,
        "skew" : 0,
        "structureFlared" : 0,
        "trafficSafetyFeatures" : {
```

```
        "bridgeRailings" : "NA",
        "transitions" : "NA",
        "approachGuardrail" : "NA",
        "approachGuardrailEnds" : "NA"
},
"historicalSignificance" : -1,
"navigationControl" : "NA",
"navigationVeriticalClearance" : 0,
"navigationHorizontalClearance" : 0,
"strucutreOpenPostedClosed" : "NA",
"typeOfService" : {
        "typeOfServiceOnBridge" : 0,
        "typeOfServiceUnderBridge" : 0
},
"structureTypeMain" : {
        "kindOfMaterialDesign" : 0,
        "typeOfDesignConstruction" : 0
},
"structureTypeApproachSpans" : {
        "kindOMaterialDesign" : 0,
        "typeOfDesignContruction" : 0
},
"numberOfSpansInMainUnit" : 0,
"numberOfApproachSpans" : 0,
"InventoryRteTotalHorzClearance" : 0,
"lengthOfMaximumSpan" : 0,
"structureLength" : 0,
"curbSidewalk Width" : {
        "leftCurbSidewalkWidth" : 0,
        "rightCurbSidewalkWidth" : 0
},
"bridgeRoadwayWithCurbToCurb" : 0,
"deckWidthOutToOut" : 0,
"minVertClearOverBridgeRoadway" : 0,
"minimumVeriticalUnderclearance" : {
        "referenceFeature" : "NA",
        "minimumVeriticalUnderclearance" : -1
},
"minLateralUnderclearOnRight" : {
        "referenceFeature" : "NA",
        "minimumLateralUnderclearance" : -1
},
"minLateralUnderclearOnLeft" : -1,
"deck" : "NA",
"superstructure" : "NA",
"substructure" : "NA",
"channelChannelProtection" : "NA",
"culverts" : "NA",
"methodUsedToDetermineOperatingRating" : -1,
"operatingRating" : 0,
"methodUsedToDetermineInventoryRating" : -1,
"inventoryRating" : 0,
"structuralEvaluation" : "NA",
"deckGeometry" : "NA",
"underclearVerticalHorizontal" : "N",
"bridgePosting" : -1,
"waterwayAdequacy" : "NA",
"approachRoadwayAlignment" : "NA",
"typeOfWork" : {
        "typeOfWorkProposed" : -1,
        "WorkDoneBy" : "NA"
},
"lengthOfStructureImprovement" : 0,
"inspectionDate" : -1,
"designatedInspectionFrequency" : -1,
"criticalFeatureInspection" : {
        "fractureCriticalDetails" : "NA",
        "underwaterInspection" : "NA",
        "otherSpecialInspection" : "NA"
},
"criticalFeatureInspectionDates" : {
```

```
                "fractureCiritcalDetailsDate" : "NA",
                "underwaterInspectionDate" : "NA",
                "OtherSpecialInspectionDate" : "NA"
        },
        "bridgeImprovementCost" : 0,
        "roadwayImprovementCost" : 0,
        "totalProjectCost" : 0,
        "yearOfImprovementCost" : 2000,
        "borderBridge" : {
                "neighboringStateCode" : "NA",
                "percentReponsibility" : -1
        },
        "borderBridgeStructureNumber" : "NA",
        "STRAHNETHighwayDesignation" : -1,
        "parallelStructureDesignation" : "NA",
        "directionOfTraffic" : -1,
        "temporaryStructureDesignation" : "NA",
        "highwaySystemOfInventoryRoute" : -1,
        "federalLandsHighways" : -1,
        "yearReconstructed" : 0,
        "deckStructureType" : "NA",
        "wearingSurface/ProtectiveSystem" : {
                "typeOfWearingSurface" : "NA",
                "typeOfMembrane" : "NA",
                "deckProtection" : "NA"
        },
        "avgDailyTruckTraffic" : -1,
        "designatedNationalNetwork" : -1,
        "pier/abutmentProtection" : -1,
        "nbisBridgeLength" : "NA",
        "scourCriticalBridges" : "NA",
        "futureAvgDailyTraffic" : 0,
        "yearOfFutureAvgDailyTraffic" : 2000,
        "minimumNavigationVerticalClearanceVerticalLiftBridge" : 0,
        "federalAgencyIndicator" : "N",
        "dateLastUpdate" : "NA",
        "typeLastUpdate" : "NA",
        "deductCode" : "Z",
        "status with 10 year rule" : "N",
        "sufficiencyRatingAsteriskField" : "NA",
        "sufficiencyRating" : -1,
        "loc" : {
                "type" : "Point",
                "coordinates" : [
                        0,
                        0
                ]
        }
}
```


Jupyter Notebook

http://faculty.ist.unomaha.edu/rgandhi/r/mongoNBI.html
Dependencies: pymongo, pandas, gridfs (image data only)

Apache Spark

Mongo is only useful for a certain kind of data

Metadata record

Do we host a record describing this dataset

Data citation/identifiers

Is this a different version

How do we deal with access? In this case, there's nothing to worry about with this dataset, but in the future.

Can someone still use this in 5-10 years

How do we upgrade Mongo,

This is active data vs a live database

Metadata via Globus?

Need to point to FHWA and github repo

Github repo needs to be tagged/versioned.